Lecture Notes Section 6

Jingyi Cui

March 4, 2021

Outline

Section 6 - Benefits of depth

- Step 1: Bounding oscillations in ReLU networks
- Step 2: constructing a deep function with many regular pieces
- Step 3: depth separation proof via region counting
- Other depth separations

Section 6 - Benefit of depth

- All preceding approximation results were about shallow networks; next, we'll discuss situations where depth helps.

- Ideally, we'd explain approximation benefits of deep networks commonly appearing in practice (e.g., those whose weights are chosen by gradient descent). We will not achieve this here, we will just give an explicit construction. Hopefully, we are not too far off from the closest analog with practical weights.

Section 6 - Benefit of depth

Plan for a first result:

- **We'll show:** there exists a function with $\mathcal{O}(L^2)$ layers and width 2 which requires width $\mathcal{O}(2^L/L)$ to approximate with $\mathcal{O}(L)$ layers.

-**Note:** it'll suffice to stick to one dimension! We'll consider higher dimension later.

- **Intuition.**

-A network with a single ReLU layer is a linear combination of some basis, meaning we are just adding things together. Thus "complexity" scales linearly with number of basis elements.

-Meanwhile, compositions can *multiplicatively* scale the "complexity." Concretely, consider the *triangle function*

$$\triangle(x) = 2\sigma_{\mathrm{r}}(x) - 2\sigma_{\mathrm{r}}(2x-1) = \begin{cases} 2x & x \leq 1/2, \\ 2-2x & x > 1/2. \end{cases}$$

We will show that the L-fold composition $\triangle^L$ has 2^L linear regions, and is hard to approximate by shallow networks.

Theorem 6.1 (Telgarsky 2015, 2016)

Theorem 6.1 (Telgarsky 2015, 2016)

For any $L \geq 2$. $f = \triangle^{L^2+2}$ is a ReLU network with $3L^2 + 6$ nodes and $2L^2 + 4$ layers, but any ReLU network g with $\leq 2^L$ nodes and $\leq L$ layers can not approximate it:

$$\int_{[0,1]} |f(x) - g(x)| \mathrm{d}x \geq \frac{1}{32}.$$

Theorem 6.1 (Telgarsky 2015, 2016)

Remark

- Previously, we used L_2 and L_∞ to state good upper bounds on approximation; for bad approximation, we want to argue there is a large region where we fail, not just a few points, and that's why we use an L_1 norm.

- To be able to argue that such a large region exists, we don't just need the hard function $f = \triangle^{L^2+2}$ to have many regions, we need them to be regularly spaced, and not bunch up. In particular, if we replaced $\triangle$ with the similar function $4x(1-x)$, then this proof would need to replace $\dfrac{1}{32}$ with something decreasing with L.

Theorem 6.1 (Telgarsky 2015, 2016)

Proof plan for Theorem 6.1 (Telgarsky 2015, 2016)
1. First we will upper bound the number of oscillations in ReLU networks. The key part of the story is that oscillations will grow polynomially in width, but *exponentially* in depth.
2. Then we will show there exists a function, realized by a slightly deeper network, which has many oscillations, which are moreover *regularly spaced*. The need for regular spacing will be clear at the end of the proof.
3. Lastly, we will use a region-counting argument to combine the preceding two facts to prove the theorem. This step would be easy for the L_∞ norm, and takes a bit more effort for the L_1 norm.

Outline

Section 6 - Benefits of depth

Step 1: Bounding oscillations in ReLU networks

Step 2: constructing a deep function with many regular pieces

Step 3: depth separation proof via region counting

Other depth separations

Step 1: Bounding oscillations in ReLU networks

Definition 6.1

For any univariate function $f\colon \mathbb{R} \to \mathbb{R}$, let $N_A(f)$ denote the number of affine pieces of f: the minimum cardinality (or ∞) of a partition of $\mathbb{R}$ so that f is affine when restricted to each piece.

Theorem 6.2

Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network with L layers of widths $(m_1, \ldots, m_L)$ with $m = \sum_i m_i$.

• Let $g\colon \mathbb{R} \to \mathbb{R}$ denote the output of some node in layer i as a function of the input. Then the number of affine pieces $N_A(g)$ satisfies

$$N_A(g) \leq 2^i \prod_{j<i} m_j.$$

$\bullet N_A(f) \leq (\frac{2m}{L})^L$

Step 1: Bounding oscillations in ReLU networks

Lemma 6.1

Let functions $f, g, (g_1, \ldots, g_k)$, and scalars $(a_1, \ldots, a_k, b)$ be given.

1. $N_A(f + g) \leq N_A(f) + N_A(g)$.
2. $N_A(\sum_i a_i g_i + b) \leq \sum_i N_A(g_i)$.
3. $N_A(f \circ g) \leq N_A(f) \cdot N_A(g)$.
4. $N_A(x \mapsto f(\sum_i a_i g_i(x) + b)) \leq N_A(f) \sum_i N_A(g_i)$.

- This immediately hints a “power of composition”: we increase the “complexity” multiplicatively rather than additively!
- It is natural and important to wonder if this exponential increase is realized in practice. Preliminary work reveals that, at least near initialization, the effective number of pieces is much smaller (Hanin and Rolnick 2019).

Step 1: Bounding oscillations in ReLU networks

Proof of Lemma 6.1.

1. Draw f and g, with vertical bars at the right boundaries of affine pieces. There are $\leq N_A(f) + N_A(g) - 1$ distinct bars, and $f + g$ is affine between each adjacent pair of bars.
2. $N_A(a_i g_i) \leq N_A(g_i)$ (equality if $a_i \neq 0$), thus induction with the preceding gives $N_A(\sum_i a_i g_i) \leq \sum_i N_A(g_i)$, and N_A doesn't change with addition of constants.
3. Let $P_A(g)$ denote the pieces of g, and fix some $U \in P_A(g)$; g is a fixed affine function along U. U is an interval, and consider the pieces of $f_{|g(U)}$; for each $T \in P_A(f_{|g(U)})$, f is affine, thus $f \circ g$ is affine (along $U \cap g_{|U}^{-1}(T)$), and the total number of pieces is

$$\sum_{U \in P_A(g)} N_A(f_{|g(U)}) \leq \sum_{U \in P_A(g)} N_A(f) \leq N_A(g) \cdot N_A(f).$$

4. Combine the preceding two. □

Step 1: Bounding oscillations in ReLU networks

Remark

The composition rule is hard to make tight: the image of each piece of g must hit all intervals of f! This is part of the motivation for the triangle function, which essentially meets this bound with every composition:

$$\triangle(x) = 2\sigma_{\mathrm{r}}(x) - 2\sigma_{\mathrm{r}}(2x-1) = \begin{cases} 2x & x \leq 1/2, \\ 2-2x & x > 1/2. \end{cases}$$

Step 1: Bounding oscillations in ReLU networks

Proof of Theorem 6.2.

To prove the second from the first, $N_A(f) \leq 2^L \prod_{j \leq L} m_j$,

$$\prod_{j \leq L} m_j = \exp \sum_{j \leq L} \ln m_j = \exp L \sum_{j \leq L} \frac{1}{L} \ln m_j \leq L \ln \frac{1}{L} \sum_{j \leq L} m_j = (\frac{m}{L})^L.$$

For the first, proceed by induction on layers. Base case: layer 0 mapping the data with identity, thus $N_A(g) = 1$. For the inductive step, given g in layer $i + 1$ which takes $(g_1, \ldots, g_{m_i})$ from the previous layer as input,

$$N_A(g) = N_A(\sigma(b + \sum_j a_j g_j)) \leq 2 \sum_{j=1}^{m_i} N_A(g_j)$$

$$\leq 2 \sum_{j=1}^{m_i} 2^i \prod_{k<i} m_k = 2^{i+1} m_i \cdot \prod_{k<i} m_k.$$

□

Outline

Section 6 - Benefits of depth

- Step 1: Bounding oscillations in ReLU networks
- Step 2: constructing a deep function with many regular pieces
- Step 3: depth separation proof via region counting
- Other depth separations

Step 2: constructing a deep function with many regular pieces

Next, let's prove that $\triangle^L$ indeed has many pieces, and regular structure; specifically, L-fold composition gives 2^{L-1} copies. Let $\langle x \rangle = x - \lfloor x \rfloor$ denote fractional part.

Lemma 6.2

$\triangle^L(x) = \triangle(\langle 2^{L-1}x \rangle) = \triangle(2^{L-1}x - \lfloor 2^{L-1}x \rfloor)$.

Step 2: constructing a deep function with many regular pieces

Proof of Lemma 6.2.

Proof by induction on $L = i$.
For base case $i = 1$, directly $\triangle^1(x) = \triangle(x) = \triangle(\langle x \rangle) = \triangle(\langle 2^i x \rangle)$.
For the inductive step, consider $\triangle^{i+1}(x)$.
If $x \in [0, 1/2]$,

$$\triangle^{i+1}(x) = \triangle^i(\triangle(x)) = \triangle^i(2x) = \triangle(\langle 2^{i-1} 2x \rangle) = \triangle(\langle 2^i x \rangle).$$

If $x \in (1/2, 1]$,

$$\begin{aligned}\triangle^{i+1}(x) &= \triangle^i(\triangle(x)) = \triangle^i(2 - 2x) \\ &= \triangle^{i-1}(\triangle(2 - 2x)) = \triangle^{i-1}(\triangle(1 - (2 - 2x))) = \triangle^i(2x - 1) \\ &= \triangle(\langle 2^i x - 2^{i-1} \rangle) = \triangle(\langle 2^i x \rangle)\ .\end{aligned}$$

□

Outline

Section 6 - Benefits of depth

Step 1: Bounding oscillations in ReLU networks

Step 2: constructing a deep function with many regular pieces

Step 3: depth separation proof via region counting

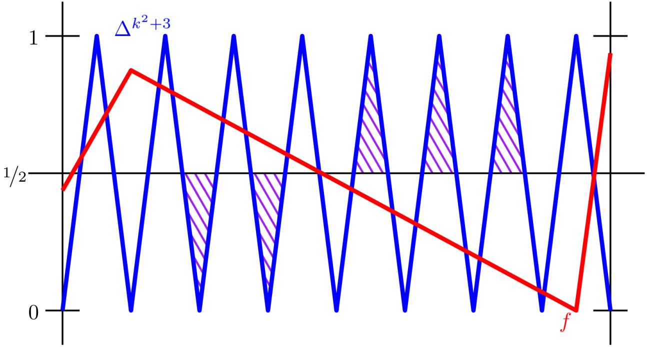
Other depth separations

Step 3: depth separation proof via region counting

We're now ready to prove our main theorem, that $\triangle^{L^2+2}$ can not be approximated by shallow networks, unless they have exponential size.

Proof of Theorem 6.1 (Telgarsky 2015, 2016).

The proof proceeds by "counting triangles."



□

Step 3: depth separation proof via region counting

Proof of Theorem 6.1 (Telgarsky 2015, 2016).

- Draw the line $x \mapsto 1/2$ (as in the figure). The "triangles" are formed by seeing how this line intersects $f = \triangle^{L^2+2}$. There are 2^{L^2+1} copies of $\triangle$, which means $2^{L^2+2} - 1$ (half-)triangles since we get two (half-)triangles for each $\triangle$ but one is lost on the boundary of $[0,1]$. Each (half-)triangle has area

$$\frac{1}{4} \cdot \frac{1}{2^{L^2+2}} = 2^{-L^2-4}.$$

- We will keep track of when g passes above and below this line; when it is above, we will count the triangles below; when it is below, we'll count the triangles above. Summing the area of these triangles forms a lower bound on $\int_{[0,1]} |f - g|$.

□

Step 3: depth separation proof via region counting

Proof of Theorem 6.1 (Telgarsky 2015, 2016).

- Using the earlier lemma, g has $N_A(g) \le (2 \cdot 2^L/L)^L \le 2^{L^2}$
- For each piece, we shouldn't count the triangles at its right endpoint, or if it crosses the line, and we also need to divide by two since we're only counting triangles on one side; together

$$\begin{aligned}
\int_{[0,1]} |f - g| &\ge [\text{number surviving triangles}] \cdot [\text{area of triangle}] \\
&\ge \frac{1}{2}[2^{L^2+2} - 1 - 2 \cdot 2^{L^2}]\ [2^{-L^2-4}] \\
&= \frac{1}{2}[2^{L^2+1} - 1]\ [2^{-L^2-4}] \\
&\ge \frac{1}{32}.
\end{aligned}$$

□

Outline

Section 6 - Benefits of depth

- Step 1: Bounding oscillations in ReLU networks
- Step 2: constructing a deep function with many regular pieces
- Step 3: depth separation proof via region counting
- Other depth separations

Other depth separations

• Our construction was univariate. Over $\mathbb{R}^d$, there exist ReLU networks with *poly*(d) notes in 2 hidden layers which can not be approximated by l-hidden-layer networks unless they have $\geq 2^d$ nodes (Eldan and Shamir 2015).
-The 2-hidden-layer function is approximately radial; we also mentioned that these functions are difficult in the Fourier material; the quantity $\int \|w\| \cdot |\hat{f}(w)| \mathrm{d}w$ is generally exponential in dimension for radial functions.
-The proof by (Eldan and Shamir 2015) is very intricate; if one adds the condition that weights have subexponential size, then a clean proof is known (Daniely 2017).

Other depth separations

-Other variants of this problem are open; indeed, there is recent evidence that separating constant depth separations is hard, in the sense of reducing to certain complexity theoretic questions (Vardi and Shamir 2020).

- A variety of works consider connections to tensor approximation and sum product networks (Cohen and Shashua 2016; Cohen, Sharir, and Shashua 2016).